Procesamiento del lenguaje Natural, nº 41 (2008), pp. 97-104 recibido 7-05-2008; aceptado 16-06-2008

MIDAS: An Information-Extraction Approach to Medical Text Classification

MIDAS: Un enfoque de extracción de información para la clasificación de texto médico

Anastasia Sotelsek-Margalef
Universidad Carlos III de Madrid
Departamento de Ingeniería Telemática
Av. de la Universidad 30, Leganés (Spain)
100025072@alumnos.uc3m.es

Julio Villena-Román
DAEDALUS, S.A.
Av. de la Albufera 321, Madrid (Spain)
jvillena@daedalus.es
Universidad Carlos III de Madrid
Departamento de Ingeniería Telemática
Av. de la Universidad 30, Leganés (Spain)
jvillena@it.uc3m.es

Resumen: Este artículo realiza una descripción de MIDAS (Medical Diagnosis Assistant), un sistema experto avanzado capaz de proporcionar un diagnóstico médico a partir de los informes radiológicos/patológicos del paciente, basado en extracción de información y aprendizaje automático a partir de historias clínicas de pacientes diagnosticados anteriormente. MIDAS fue diseñado para participar en la competición Medical Natural Language Processing Challenge 2007. Específicamente, el sistema automatiza la asignación de códigos ICD-9-CM (International Classification of Diseases) a informes medicos, logrando unos buenos resultados de precisión.
Palabras clave: sistema experto, diagnóstico, texto médico, lenguaje natural, extracción de información, clasificación automática, códigos ICD-9-CM.

Abstract: This article describes MIDAS, an advanced expert system that is able to suggest medical diagnosis from the radiological/clinical patient records, based on information extraction and machine learning from clinical histories of previously diagnosed patients. MIDAS was designed to participate in the 2007 Medical Natural Language Processing Challenge. Specifically, it automates the assignment of ICD-9-CM codes to radiology reports, achieving good precision rates.
Keywords: Expert system, medical diagnosis, medical text, natural language, information extraction, automatic classification, ICD-9-CM codes.

1 Introduction

The fact that clinical information systems can improve medical care and reduce health costs has been in the academic agenda for quite some time. Nonetheless, nowadays patient data is still stored in narrative form by many hospitals, which produces a great quantity of information that, beyond the clinical visit, has limited utility because of its high volume and poor accessibility. However, attempts to address the problem of free text processing have led to demand for software that simulates and complements what people are able to do.

This article describes MIDAS (Medical Diagnosis Assistant), an advanced expert system that is able to suggest medical diagnosis from the radiological/clinical patient records, based on information extraction and machine learning from clinical histories of previously diagnosed patients. For this task, free text is turned into actionable knowledge using Natural Language Processing (NLP) techniques which is then used to train machine-learning systems to perform clinical free text classification.

MIDAS was specifically designed to participate in the 2007 Medical Natural Language Processing Challenge (CMC, 2007),

ISSN: 1135-5948

an international challenge task on the automated processing of clinical free text, hosted by the Computational Medicine Center, a collaborative medical research centre between Cincinnati Children's Hospital Medical Center and the University of Cincinnati Medical Center).

MIDAS can be considered as one of the latest successors of MYCIN, the first expert system in history developed in the early 1970s at Stanford University, which was designed to diagnose infectious blood diseases (Shortliffe 1976).

2 Background and related work

The task of classifying physicians' diagnoses has been previously done. Gundersen et al. (1996) presented a system designed to assign diagnostic ICD-9-CM codes to the free text of admission diagnoses. This system encoded the diagnoses using categories from a standard classification scheme based on a text parsing technique informed with semantic information derived from a Bayesian network.

Yang et al. developed ExpNet (Yang, 1994), which comprised a machine learning method for automatic coding of medical diagnoses. This system offered improvements in scalability and computational training efficiency using Linear Least Squares Fit and Latent Semantic Indexing. Pakhomov et al. (1996) scaled up this groundwork with a hybrid approach consisting of example based classification and a simple but robust classification algorithm (naive Bayes) in order to improve the efficiency of diagnostic coding.

Other machine learning algorithms have been used to investigate classification problems related to medical reports. These include decision trees (Johnson, 2002), maximum entropy and symbolic rule induction (Nigam, 1999) among others.

As far as information extraction goes, many systems utilize patterns for extraction. Earlier pattern-based work such as AutoSlog (Riloff, 1993) solved the problem of domain specific dictionaries by developing a system that automatically builds domain specific dictionaries of concepts by extracting information from text. Other systems such as MedLEE (Friedman, 1994) used patterns to represent particular scenarios or events where the desired information is found by mapping clinical information into a structured representation containing clinical terms.

Linguistic variations of existing patterns have also been explored to increase domain patterns (Hobbs, 2003). These types of systems have the advantage of being able to "learn" patterns without the need of massive amounts of hand-tagged training data. Other groups have worked on the problem of automated biomedical concept recognition. The SAPHIRE system designed by Hersh et al. (1995) automatically encodes UMLS concepts using lexical mapping. The lexical approach is computationally fast and useful for real-time applications. More recently, Zou et al. (2003) developed IndexFinder to add syntactic and semantic filtering to improve performance on top of lexical mapping.

3 Description of Data

The data provided in the framework of the Computational Medicine Center's 2007 Medical Natural Language Processing Challenge (CMC, 2007) was used. The corpus was collected from the Cincinnati Children's Hospital and included a repertoire of codes covering a substantial proportion of actual paediatric radiology activity. It was initially developed to train machine learning systems dedicated to automatic billing of medical records and other related activity. The set sample developed is representative of the problem: it has enough data in the well-represented classes for the automatic labeller to perform adequately and provides a proportionate representation of low-frequency classes.

An ICD-9-CM *(International Classification of Diseases, 9th Revision, Clinical Modification)* code is a 3 to 5 digit number with a decimal point after the third digit. Codes are organized in a hierarchy, with the highest levels of the hierarchy lumping codes together by assigning consecutive numbers, e.g.:

(580-629) GENITOURINARY SYSTEM
-(580-589) NEPHRITIS AND NEPHROSIS
- 580 Acute glomerulonephritis
-580.8 Other specified pathological lesion in kidney
-580.81 Acute glomerulonephritis in diseases classified elsewhere
-580.89 Other

Two sections in a radiology report are fundamental for assigning ICD-9-CM codes: clinical history, provided by an ordering physician before a radiological procedure, and impression, reported by a radiologist after the

procedure. The language of clinicians is fundamental to patient care, but lacks the structure and clarity necessary for natural language analysis. These clinical annotations are dense with medical jargon and acronyms that often have multiple meanings. To resolve the ambiguities found in the free text, a series of clinical disambiguation rules were developed using clinical experts to translate the ambiguous terms, clinical acronyms, and abbreviations.

Finally, the data was converted to XML with two top-level subdivisions: texts and codes. Figure 1 shows a fragment of the patient record file.

```
<doc id="97636670" type="RADIOLOGY_REPORT">
 <codes>
  <code type="ICD-9-CM">786.2</code>
 </codes>
 <texts>
  <text type="CLINICAL_HISTORY">Eleven year old with ALL, bone marrow transplant on Jan. 2, now with three day history of cough.</text>
  <text type="IMPRESSION">1. No focal pneumonia. Likely chronic changes at the left lung base. 2. Mild anterior wedging of the thoracic vertebral bodies.</text>
 </texts>
</doc>
<doc id="99636934" type="RADIOLOGY_REPORT">
 <codes>
  <code type="ICD-9-CM">593.70</code>
  <code type="ICD-9-CM">599.0</code>
 </codes>
 <texts>
  <text type="CLINICAL_HISTORY">10-year 5-month - old female with history of urinary tract infection. Patient had nuclear cystogram and was found to have left grade II vesicoureteral reflux. Last ultrasound of Jan. 27, 2001 demonstrated little growth of the right kidney compared to the left, otherwise stable renal ultrasound.</text>
  <text type="IMPRESSION">1. Normal renal ultrasound with interval growth of the both kidneys.</text>
 </texts>
</doc>
```

Figure 1: Example of patient data

4 System Architecture

The system is designed according to a modular cascade architecture (Figure 2). The first module extracts and structures clinical information from textual radiology reports and translates the information to terms in a controlled vocabulary so that clinical information can be accessed by further automated procedures. The objective is to automate sufficient understanding of clinical records contents to be able to label all the phrases in them that contained information related to symptoms and signs of diseases that would be later used in the training of the classification algorithm. Each symptom, not always composed of a single word, was labelled as *present*, *absent*, *family*, *history*, *past* or *unknown* following a set of linguistic context rules.

The information extraction task is based on semantic pattern matching allowing for the identification of particular values of interest which are embedded within free text and determining a given value's categorization. Keyword extraction from the free-text reports is susceptible to all the problems that result from the complexities of natural language, such as grammatical ambiguities, synonymy, negation of concepts and distribution of concepts (Sager, 1997).

Finally, a classifier is built based on a suitable ML algorithm. Weka (Witten, 2005) was used for the experiments.

4.1 Linguistic Preprocessor

Clinical documents usually contain syntactic structures that are generally considered incorrect. Shorthand and telegraphic writing styles are common in radiology reports (in both fields). In addition, syntactic tagging implies that every word or phrase must be tagged whereas in our case only the targeted information needs to be identified. Sentences that are irrelevant to the domain can be effectively ignored without affecting the final classification. Therefore no syntactic parser was used in our system.

The first step to translate all relevant information into structured form is to standardize the character representation of the text and remove custom text formatting. Simple heuristic rules eliminate or modify line feeds, sequences of blanks between words and punctuation marks.

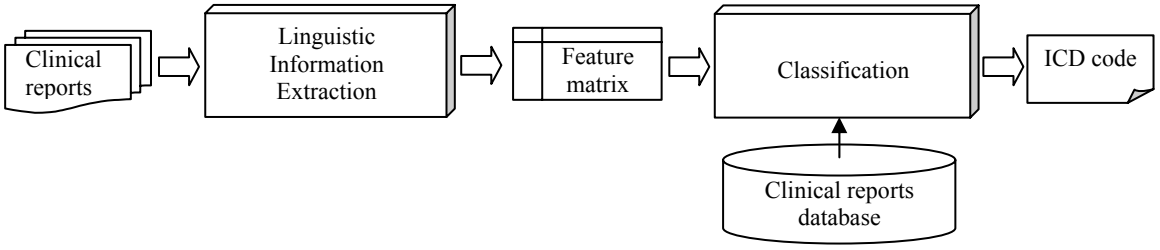

Figure 2: Overview of System Architecture

Then the structural analyzer segments the report into sections (e.g., clinical history and impression), sentences and words. Stop words are filtered based on their level of usefulness within this context and according to their usage. Words as *also* and *or* are eliminated since they are not useful in the labelling process.

The lexicon was manually developed. Both single words and multiple words phrases (multiword units) were included. Multiword combinations provide better retrieval performance allowing for a better capture of the content of the documents. In addition, abbreviations, proper names and descriptive adjectives that may not be found in electronic medical glossaries have been also considered.

A lexical lookup to identify multi-word phrases is performed. For instance, the sentence *history of pneumonia* would be considered a sequence of two terms, *history of* and *pneumonia* because the first term is considered a multiword phrase in the lexicon. In the next stages, these multiword phrases are treated as single entities.

The next phase of the process consists on the mapping of different forms of the same words and multiword units into a (single) term within the controlled vocabulary lexicon. In other words, a synonym knowledge base that consists of standard forms and their corresponding synonyms is used. If any value matches the argument of a synonym entry in the synonym knowledge base, it is substituted for the controlled vocabulary concept.

While the system was designed to consider every reference made to the symptoms, phrases like *rule out*, *evaluate for* or *look for* do not appear to be useful for classification since in the same report there is another reference to the sign or symptom indicating its diagnosis (e.g., *no findings consistent with acute pneumonia*). To address this issue, these phrases are eliminated in the pre-processing stage without causing a loss of relevant information.

4.2 Structured Representation

Our data model can be described as a set of attributes (e.g., signs and symptoms) with their corresponding values. Our objective is to extract information on the existence and diagnostic interpretation of findings. Looking up isolated word meanings is not enough to make distinctions on whether the symptom is *present*, *absent*, or *not mentioned* at all. Furthermore other tags such as *family* were added to avoid misinterpretation of the presence of a symptom in a patient when, for example, the symptom actually was suffered by a sibling.

Rules, specific to the writing style of medical reports, were used to assign the different tags. Negation, a particularly troublesome aspect of natural language processing, is specified as an atomic category *absent*. The target structure for negation is a finding qualifier whose value is part of a list of key words provided (e.g., *no*, *without*).

Since each attribute may have more than one possible value associated to it, there is a need to determine the value which best corresponds with the attribute. To resolve inconsistencies when labelling the attributes, there is an order in which they are looked for in the text. The label *absent* has more priority than *history*. In the sentence *no history of pneumonia*, the attribute *pneumonia* is therefore correctly labelled as *absent*.

For multi-valued attributes such as the age of the patients, regular expressions are used. Regular expressions have been widely used for lexical pattern matching tasks. Each attribute is assigned a set of regular expressions which represent every possible way a valid value for that attribute can be lexically expressed within a document (Meng, 2004). The label *suspected* is associated with certainty information related to the finding. Semantic relations such as *could represent*, *suggesting* and *consistent with* are recognized and the finding to which they are referring to is assigned this label.

Because there are many words and phrases linked to this type of information, and because their underlying meanings are vague, they all are mapped into one category only. We considered extracting more detailed information in terms of low, moderate or high certainty but we finally rejected that idea. In other applications, handling qualitative information more precisely may be important, in which case more labels could be desirable.

Parallel findings, such as *hyperinflated lungs without pleural effusion* are represented as independent findings, the first labelled with the tag *present* and the second with the tag *absent*. In the case of sentences containing *or* and *and*, such as *no pneumonia or atelectasis* and *history of cough and fever*, the interpretation made consists of two findings. For the former case both *pneumonia* and *atelectasis* are labelled as *absent*, for the latter *cough* and *fever* are labelled as *history*.

4.3 Classification

The classifier was built using Weka, a suite of machine learning software that implements numerous machine learning algorithms. The first problem that we encountered was how to handle a multi-labelled data set, as Weka does not support multi-labelled learning. The chosen solution was to create new artificial classes corresponding to the combination of labels (e.g., *780.6-786.2*).

Several algorithms were evaluated, but after the preliminary evaluations, two of them were finally selected: the classical C4.5 decision tree algorithm (Quinlan, 1993), namely J48 in Weka, and the k-Nearest-Neighbour classifier (Mitchell, 1997), IBk in Weka.

5 Evaluation

The provided 1,954 patient reports contained 29 different ICD-9-CM labels (e.g. *780.6*) that formed 89 distinct combinations (e.g. the combination *780.6-786.2*).

Code	Description	No.
786.2	Cough	155
599.0	Urinary tract infection	114
593.70	Unspecified or w/o reflux nephropathy	80
780.6-786.2	Fever-Cough	76
486	Pneumonia	66
780.6	Fever	41
591	Hydronephrosis	40
786.50	Chest pain	32
596.54	Neurogenic bladder	31
788.30	Urinary incontinence	29
599.7	Hematuria	25
786.07	Wheezing	24
795.5	Nonspecific reaction to tuberculin test w/o tuberculosis	16
591-593.89	Hydronephrosis-disorders of kidney and ureter	16
493.90	Asthma	15
277.00	Cystic Fibrosis	15
518.0	Pulmonary collapse	12
786.07-786.2	Wheezing-Cough	12
759.89	Congenital malformation	11
596.54-741.90	Neurogenic bladder-w/o hydrocephalus	11

Table 1: Distribution of radiology reports in the largest categories of the training set.

Table 1 shows the number of reports per category, the ICD-9-CM code and its description for those categories with more than 10 reports in the training set.

Three different experiments were performed, one based on J48 (decision trees) and the other two based on IBk (kNN), using two values for k (number of neighbours). Experiments were run using a 10-fold cross validation test. Results are shown in Table 2. The standard evaluation metric of F-Measure, the weighted harmonic mean of precision and recall, was calculated, using the micro-averaged figure (value is first calculated for each category and then averaged). J48 achieves the best performance.

ML algorithm	(micro-averaged) F-Measure
J48	0.8004
IBk (k=1)	0.7671
IBk (k=2)	0.7625

Table 2: F-Measure values

Table 3 shows the detailed accuracy per class of J48 algorithm. Notice that precision and recall are significantly better for those categories with a high number of instances (shown in Table 1).

Code	Precision	Recall	F-Measure
786.2	0.913	0.91	0.911
599.0	0.872	0.93	0.9
593.70	0.84	0.882	0.861
780.6-786.2	0.842	0.954	0.894
486	0.841	0.879	0.859
780.6	0.837	0.878	0.857
591	0.729	0.765	0.747
786.50	0.87	0.923	0.896
596.54	0.757	0.903	0.824
788.30	0.94	0.81	0.87
599.7	0.796	0.86	0.827
786.07	0.816	0.833	0.825
795.5	0.875	0.875	0.875
591-593.89	0.92	0.719	0.807
493.90	0.762	0.533	0.627
277.00	1	1	1
518.0	0.625	0.4	0.488
786.07-786.2	0.75	0.875	0.808
759.89	1	0.818	0.9
596.54-741.90	0.615	0.364	0.457

Table 3: Detailed accuracy per class.

If categories with less than 5 reports are filtered out from data, the percentage of correctly classified instances is noticeably higher (Table 4).

Algorithm	F-Measure	Increment
J48	0.8586	7.3%
IBk (k=1)	0.8255	7.6%
IBk(k=2)	0.7959	4.3%

Table 4: Results for categories with 5 or more instances.

Regretfully we were not able to submit any experiment to the challenge, due to delays during the system development. In fact, only 44 out of the over 120 registered participants in the challenge finally submitted their results. The best and worst systems achieved F-Measure values of 0.8908 and 0.1541, respectively. The average value was 0.7670 with a standard deviation of 0.1340. In addition, 21 systems get F-measure values between 0.81 y 0.90.

The groups in 1st and 3rd position used machine learning approaches, whereas the system in 2nd position was based on symbolic methods. Actually the best system was based on a particular implementation of C4.5 algorithm, the same as our system.

6 Conclusions and Future Work

The expected potential of such systems is to make available a large body of clinical information that would otherwise be inaccessible for applications other than manual physician review. We do not intend to replace coded data entry, but we offer a solution for the virtual enrolment of previously evaluated patients that would benefit research studies, teaching hospitals and physicians with a large workload in emergency situations.

The accuracy and hence the utility of a medical natural language processor relies heavily on the number and diversity of high-quality training examples. Furthermore, the accuracy of a language system depends on the specific information that it extracts. The important types of information for a given type of study should be established a priori, allowing system developers to emphasize training on high-priority information items.

Natural language used within patient documents is limited in word and phrasal variation. Thus the linguistic context in which the information to be extracted resides may only take on several basic structural forms. With a reasonable amount of training, which in MIDAS means labelling domain specific symptoms, any system built with the described methodology can obtain successful results.

We believe that our system could allow medical experts, making the necessary configuration changes, to tune the processor to their particular field without possessing expertise in the technical aspects of the system.

Moreover, although MIDAS has been specifically applied to the radiology domain, the proposed methodology is modular and extensible and can be ported to other clinical domains. Explorations of the system's adaptability to new clinical domains will be further conducted.

References

Computational Medicine Center (CMC). 2007. *Medical Natural Language Processing Challenge.* http://computationalmedicine.org /challenge

Friedman C, Alderson P, Austin J, Cimino JJ and Johnson SB. 1994. A general natural language text processor for clinical radiology. *Journal of American Medical Informatics Association*, March 1994, 1(2):161-174.

Gundersen ML, Haug PJ, Pryor TA, et al. 1996. Development and evaluation of a computerized admission diagnoses encoding system. *Comp Biomed Res*; 29(5): 351–72.

Hersh WR, Hickam D. 1995. Information retrieval in medicine: the SAPHIRE experience. *Medinfo, 8* Pt 2:1433–7.

Hobbs JR. 2003. Information extraction from biomedical text. *Journal of Biomedical Informatics*.

Johnson D., et al. 2002. A decision tree based symbolic rule induction system for text categorization. *IBM Systems Journal,* 41(3).

Meng F, Chen AA, Son RY, Taira RK, Churchill BM, Kangarloo H. 2004. Information Extraction Using Semantic Patterns for Populating Clinical Data Models. *METMBS'04*: 10-16

Mitchell TM. 1997. *Machine Learning*. McGraw-Hill.

Nigam K, Lafferty J, McCullum A. 1999. Using Maximum Entropy for Text classification. In IJCAI-99 *Workshop on Machine Learning for Information Filtering*.

Pakhomov S, Buntrock J, Chute CG. 2006. Automating the assignment of diagnosis codes to pacient encounters, *Journal of American Medical Informatics Association*, 13: 516-525.

Quinlan JR. 1993. *C4.5. Programs for Machine Learning*. Morgan Kaufmann.

Riloff E. 1993. Automatically constructing a dictionary for information extraction tasks, *Proceedings of the 11th National Conference on Artificial Intelligence*, AAAI Press, pp. 811-816.

Sager N. 1997. *Medical Language Processing: Computer Management of Narrative Data.* Springer-Verlag, New York.

Shortliffe E. 1976. *MYCIN: Computer-Based Medical Consultations*. Elsevier, New York.

Witten IH, Frank E. 2005. Data Mining-Practical Machine Learning Tools and Techniques. Elsevier Inc.

Yang Y, Chute CG. 1994. An application of Expert Network to clinical classification and MEDLINE indexing. *Journal of American Medical Informatics Association* 18; 157-61.

Zou Q, Chu WW, Morioka C, et al. 2003. IndexFinder: a method of extracting key concepts from clinical texts for indexing. *Proceedings of AMIA Symposium*; 763–7.

A Appendix 1: Web interface

The web interface of the system is shown in Figure 3. There are two textboxes for writing the clinical history (physician information) and impression (radiologist report) and the diagnosis is shown in real-time after clicking on the "Diagnose" button.

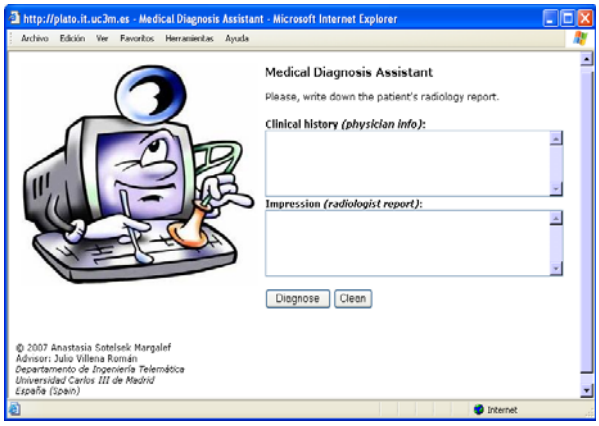

Figure 3: Web interface

B Appendix 2: List of symptoms

The list of symptoms in the MIDAS system covers the whole range of illnesses included in the CMC challenge, a substantial proportion of actual paediatric radiology activity.

abdominal pain, air space disease, anomal, anuresis, asthma, atelectasis
Beckwith Wiedemann syndrome, bronchiectasis
calculi, cardiopulmonary disease, chest pain, chest tightness, congestion, consolidation, cough, cystic fibrosis
deflux, difficulty breathing, dilatation, distended bladder, duplication
enuresis
fever, flank pain
hematuria, hemihypertrophy, horseshoe kidney, hydronephrosis, hydroureter, hydroureteronephrosis, hyperinflat, hypertrophy, hypoventilation
infiltrate, interval growth
lobe collapse, loss of appetite, lymphadenopathy
mass, myelomeningocele
neurogenic bladder, normal chest, normal heart, normal kidney, normal lungs
peribronchial cuffing, peribronchial thickening, pleural effusion, pneumonia, pneumothorax, positive PPD, post void residual, proteinuria, pyelectasis, pyelocaliectasis, pyeloplasty
reactive airway, reflux, renal transplant
shortness of breath, sore throat, spina bifida
tachypnea, tuberculosis, Turner syndrome
ureteropelvic junction obstruction, unilateral kidney, ureterocele, urinary incontinence, urinary tract infection, urothelial thickening
vesicoureteral reflux, voiding dysfunction, vomiting
wheezing, Williams syndrome, Wiskott Aldrich

C Appendix 3: List of synonyms

airway disease, reactive airway
calculi, calculus, calcifications
cough, coughing
cystic fibrosis, CF
difficulty breathing, work of breathing
disease, illness
duplication, duplicated kidney
examination, evaluation, exam, study
family, history of, siblings, brothers
fever, febrile
hyperinflation, hyperinflated lungs
interval growth, interval renal growth
may, could
normal, unremarkable, stable, clear, normal radiographic appearance of the, normal radiographs of the, normal sonographic appearance of the, normal examination of the, normal sonographic examination of the

normal heart, heart normal
normal kidney, kidney normal, normal renal
normal lungs, lungs normal
post void, postvoid
postive PPD, reactive PPD
prior, previous, past, status post, had
probable, may represent, likely representing, likely represent, probably representing, favored to represent, raising the question of, can be associated, may be related, sometimes associated with, consistent with
probable, worrisome, questionable, suggesting, suggest, suggests, suggestive, suspected, presumed, suspicion, possible, likely, unsure
radiograph, x ray
represent, reflect
shortness of breath, breathlessness
sonography, ultrasound, sonogram
tuberculosis, TB
urinary incontinence, wetting
urinary tract infection, UTI, UTIs
viral disease, viral infection
vomiting, emesis
vs, versus, is favored over, favored over